Iniziato	martedì, 15 novembre 2022, 11:22
Stato	Completato
Terminato	martedì, 15 novembre 2022, 11:23
Tempo impiegato	7 secondi
Valutazione	**0,00** su un massimo di 12,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 5,00

Un'agenzia immobiliare desidera costruire un modello predittivo per il prezzo degli immobili abitativi di una particolare città venduti sul sito nel 2019.
Per un campione di 90 immobili vengono quindi raccolte varie informazioni tra cui:

- l'ultimo prezzo listato per una proprietà sul sito (`P`, in migliaia di euro)
- la dimensione in metri quadri (`m2`)

Alcune statistiche descrittive del dataset sono fornite:

```
summary(df)
```

```
       P              m2
 Min.   :  82   Min.   : 32
 1st Qu.: 590   1st Qu.:123
 Median : 805   Median :209
 Mean   : 819   Mean   :195
 3rd Qu.:1065   3rd Qu.:272
 Max.   :1601   Max.   :329
```

In prima istanza l'analista stima un modello lineare semplice:

```
fit_M <- lm(P~m2, data = df)
summary(fit_M)
```

```

 Call:
 lm(formula = P ~ m2, data = df)

 Residuals:
    Min     1Q Median     3Q    Max
 -389.9  -87.4  -18.7  105.8  395.6

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  223.653     38.625    5.79  1.1e-07 ***
 m2             3.047      0.181   16.83  < 2e-16 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 147 on 88 degrees of freedom
 Multiple R-squared:  0.763,   Adjusted R-squared:  0.76
 F-statistic:  283 on 1 and 88 DF,  p-value: <2e-16
```

Costruisce poi un intervallo di confidenza (al 95%) per il coefficiente angolare del modello.

Usando il modello `fit_M` costruisce poi intervalli di confidenza per la stima del prezzo medio di due immobili di 100 e 500 m^2.

```
ci_100 <- predict(fit_M, newdata = data.frame(m2 = 100),
                  interval = "confidence")
ci_500 <- predict(fit_M, newdata = data.frame(m2 = 500),
                  interval = "confidence")
```

Su richiesta di un collega l'analista usa anche i dati sul prezzo delle case in centinaia euro (e non in migliaia)

```
df$prezzo <- df$P * 10
fit_dec <- lm(prezzo~m2, data = df)
```

Si indichi se il modello `fit_M` è significativo rispetto al modello nullo. Si indichi il valore del limite superiore dell'intervallo di confidenza per il coefficiente angolare. Si indichi quale degli intervalli di confidenza (`ci_100` e `ci_500`) è più ampio. Si indichi infine se il modello `fit_dec` risulta significativo rispetto al modello nullo e se ne indichi il valore dell'intercetta.

a. ○ Il modello `fit_M` è significativo rispetto al modello nullo.
○ Il modello `fit_M` non è significativo rispetto al modello nullo.
○ Non è possibile stabilire la significatività del modello `fit_M`.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: Il modello `fit_M` è significativo rispetto al modello nullo.

b. Limite superiore dell'intervallo di confidenza per il coefficiente angolare:.

✖

c. ○ Gli intervalli hanno la stessa ampiezza.
○ ci_500 è più ampio di ci_100.
○ ci_500 è meno ampio di ci_100.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: ci_500 è più ampio di ci_100.

d. ○ Il modello `fit_dec` non è significativo rispetto al modello nullo.
○ Non è possibile stabilire la significatività del modello `fit_dec`.
○ Il modello `fit_dec` è significativo rispetto al modello nullo.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: Il modello `fit_dec` è significativo rispetto al modello nullo.

e. Valore dell'intercetta del modello `fit_dec`.

✖

Guardando il valore del p-value per la F-statistic di nota che è molto piccolo: il modello è significativo rispetto al modello nullo.

L'intervallo di confidenza si può derivare con: $\hat{\beta} \pm z_{alpha/2} * se(\hat{\beta}) = 3.047 \pm 1.96 * 0.181$, quindi (2.692, 3.402)

Il valore di 500 m^2 è molto estremo rispetto al campione usato per stimare i parametri del modello, mentre il valore 100 è abbastanza centrale nel range di valori nel dataset `df`: l'intervallo di confidenza `ci_500` è più ampio di `ci_100`.

Il modello `fit_dec` è un modello che usa una semplice trasformazione lineare della variabile risposta: la significatività della relazione tra prezzo e dimensione della casa non cambia. L'intercetta indica il prezzo di una casa di dimensione 0 m^2: il valore sarà il valore dell'intercetta di `fit_M` moltiplicato per 10.

a. Vero / Falso / Falso
b. 3.41
c. Falso / Vero / Falso
d. Falso / Falso / Vero
e. 2236.53

Domanda 2

Risposta non data

Punteggio max.: 3,00

Un'azienda desidera capire se le vendite del loro prodotto nei loro punti vendita sono in qualche modo legate alle vendite del prodotto nel loro negozio digitale. Per indagare su questa relazione vengono compilate le informazioni sulle vendite settimanali nei negozi fisici (centinaia di pezzi, variabile X) e digitale (centinaia di pezzi, variabile Y). L'analista decide di fare un grafico di dispersione (scatterplot) delle due variabili e di usare una regressione lineare semplice per studiare la relazione tra X e Y.

```
fit <- lm(y~x)
```

```
summary(fit)
```

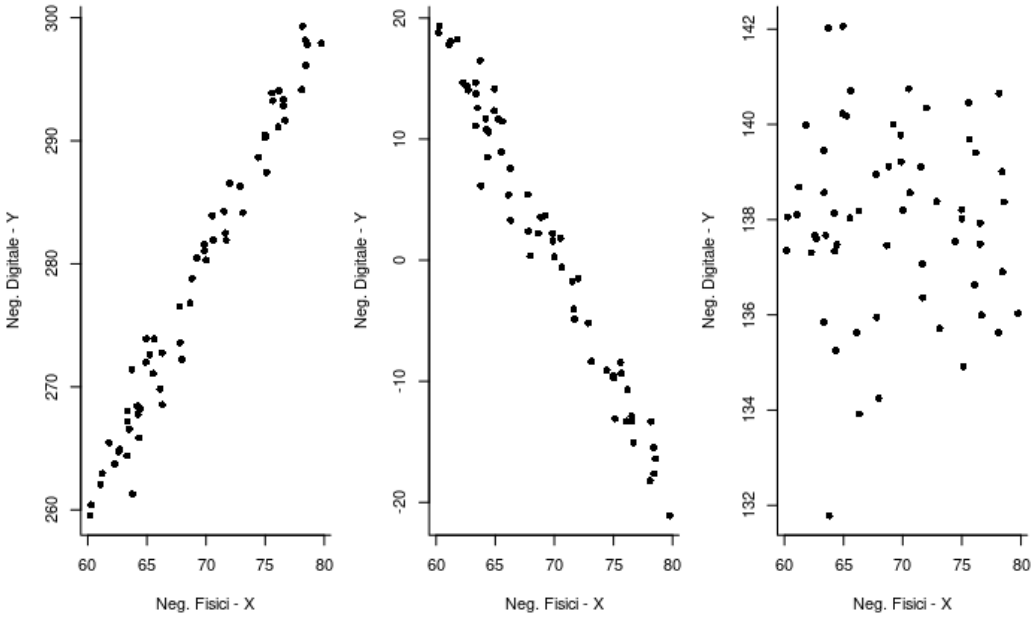
```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-6.259 -0.917  0.060  1.275  4.027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 138.6946     3.1402   44.17   <2e-16 ***
x            -0.0102     0.0452   -0.23     0.82
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.99 on 58 degrees of freedom
Multiple R-squared:  0.000875,  Adjusted R-squared:  -0.0164
F-statistic: 0.0508 on 1 and 58 DF,  p-value: 0.822
```

La Figura mostra vari grafici di dispersione: si indichi quale dei grafici è quello che mostra i dati analizzati dall'analista.



Per rendere l'interpretazione del modello più semplice l'analista decide di trasformare la variabile X, che registra le centinaia di mezzi venduti, nella scala originale di pezzi venduti. Dopo questa trasformazione stima di nuovo un modello di regressione semplice.

```
xt <- 100*x
fit100 <- lm(y ~ xt)
```

Si indichino le caratteristiche di alcune proprietà del modello `fit100`.

a. I dati usati sono mostrati nel pannello A.

I dati usati sono mostrati nel pannello B.

I dati usati sono mostrati nel pannello C.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: I dati usati sono mostrati nel pannello C.

b. Si indichi il valore del p-value per il test di significatività del coefficiente angolare stimato del modello fit100.

✖

c. Si indichi il valore dell'intercetta stimato nel modello fit100.

✖

Il coefficiente angolare stimato è negativo ma non fortemente significativo: si deve quindi cercare un grafico in cui Y diminuisca al crescere di X, ma con una relazione non molto marcata. Il grafico che mostra i dati analizzati è quindi il grafico C.

Il modello originale assume che:

$$E[Y|X = x_i] = \beta_0 + \beta_1 * x_i$$

il modello in cui si usa $XT = 100 * X$ assume che:

$$E[Y|XT = xt_i] = \tilde{\beta}_0 + \tilde{\beta}_1 * xt_i$$

questo si può riscrivere come:

$$E[Y|XT = xt_i] = \tilde{\beta}_0 + \tilde{\beta}_1 * 100 * x_i$$

quindi $\tilde{\beta}_1 * 100 = \beta_1$, mentre l'intercetta dei due modelli è uguale: l'intercetta rappresenta il valore atteso di Y quando il predittore è 0. Infine la significatività della relazione tra X ed Y o XT e Y è la stessa.

a. Falso / Falso / Vero
b. 0.82
c. 138.69

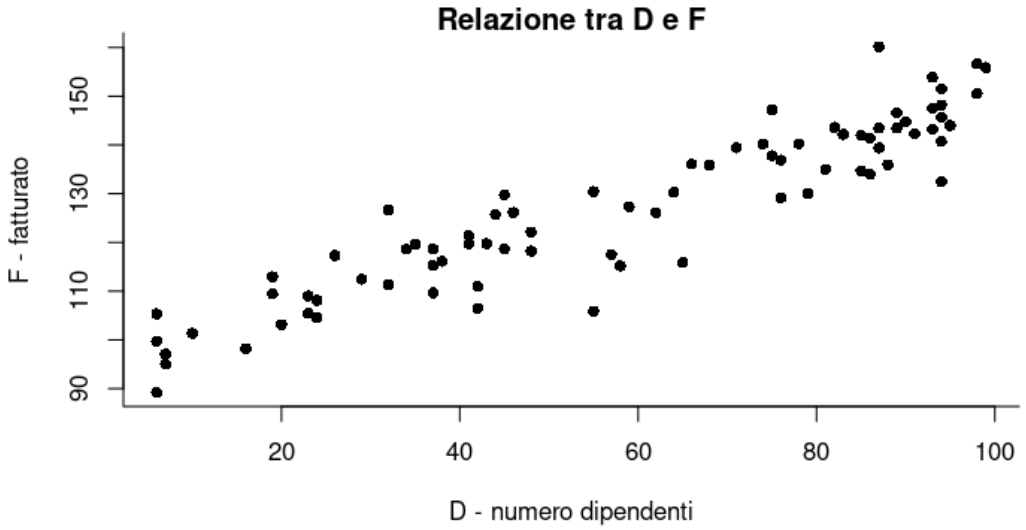
Domanda 3

Risposta non data

Punteggio max.: 4,00

L'ufficio commerciale di una catena di supermercati desidera indagare la relazione tra il numero di dipendenti e il fatturato nei diversi punti vendita della catena. Per un campione di 80 punti vendita vengono quindi raccolte informazioni per l'anno passato sul numero di dipendenti (indicato con la variabile D) e il fatturato medio mensile in centinaia di migliaia di euro (indicato con la variabile F).

La relazione tra F e D nel campione analizzato è mostrata nel grafico:



In prima istanza l'analista stima un modello lineare semplice:

```
fit <- lm(f~d, data = df)
```

che confronta contro il modello nullo tramite un test ANOVA:

```
fitNull <- lm(f~1, data = df)
anova(fitNull, fit)
```

```
 Analysis of Variance Table

 Model 1: f ~ 1
 Model 2: f ~ d
   Res.Df   RSS Df Sum of Sq   F Pr(>F)
 1     79 22675
 2     78  3213  1     19462 472 <2e-16 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calcola poi intervalli di confidenza per due valori del predittore: $d = 60$ e $d = 110$.

```
ci60 <- predict(fit, newdata = data.frame(d=60), interval = "confidence")
ci110 <- predict(fit, newdata = data.frame(d=110), interval = "confidence")
```

Infine desidera stimare un modello che prenda in considerazione la posizione del punto vendita (P) che è codificata in una variabile a tre livelli:

```
table(df$p)
```

```
       altro centro_citta    periferia
          24           22           34
```

```
fitLoc <- lm(f~d+p,data=df)
```

```
fitLoc$df.residual
```

```
[1] NA
```

Si indichi quale equazione è quella che descrive il modello stimato usando i dati mostrati nella figura. Si indichi inoltre quale dei due intervalli di confidenza è più ampio. Si indichi poi il numero dei gradi di libertà residui per il modello `fitLoc`. Infine si indichi se un modello

```
fitInv <- lm(d~f, data = df)
```

risulterebbe significativo (usando un livello di significatività $\alpha = 0.05$).

a. Il modello stimato è: F = 92 - 0.56 D.
Il modello stimato è: F = 95 + 0.54 D.
Il modello stimato è: F = 92 + 7.6 D.
Il modello stimato è: F = 4.1 + 0.54 D.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: Il modello stimato è: F = 95 + 0.54 D.

b. ci60 è meno ampio di ci110.
ci60 è più ampio di ci110.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: ci60 è meno ampio di ci110.

c. Si indichi il numero di gradi di libertà residui per il modello fitLoc (mancante nell'output):.

✖

d. Il modello `fitInv` è significativo.
Il modello `fitInv` non è significativo.

Punteggio ottenuto 0,00 su 1,00

La risposta corretta è: Il modello `fitInv` è significativo.

La figura mostra una relazione positiva e in cui estrapolando al caso in cui un punto vendita abbia 0 dipendenti ci sarebbe comunque un cospicuo fatturato. Si osserva poi che il cambiamento da 0 a 100 dipendenti porta ad aumenti del fatturato nell'ordine di decine di migliaia di euro. L'unica opzione che soddisfa tutti i requisiti è

Il modello stimato è: F = 95 + 0.54 D

Il valore di 60 dipendenti è più prossimo al valore medio dei dipendenti del valore 110: gli intervalli di confidenza sono più ampi tanto più sono valutati per valori del predittore distanti dalla media. Si ha quindi che ci60 è meno ampio di ci110.

Il modello `fitLoc` usa 4 gradi di libertà e ha quindi un totale di 80 - 4 gradi di libertà.

Dato che il modello `fit` è significativo `fitInv` dovrà anch'esso essere significativo.

a. Falso / Vero / Falso / Falso
b. Vero / Falso
c. 76.00
d. Vero / Falso

◄ Esercizi 1 - Regressione Lineare Semplice

Vai a...

Quiz Modelli Regressione Multipla ►